

The Indra Representation Hypothesis — Supplementary Material

Jianglin Lu^{1*} Hailing Wang^{1*} Kuo Yang¹ Yitian Zhang¹ Simon Jenni³ Yun Fu^{1,2}

¹Department of Electrical and Computer Engineering, Northeastern University

²Khoury College of Computer Science, Northeastern University

³Adobe Research

A Proof of Theorem

Before proceeding with the proof, we introduce the following definition and lemma.

Definition 1 (\mathcal{V} -functor). A \mathcal{V} -functor $P : \mathcal{D} \rightarrow \mathcal{E}$ between \mathcal{V} -categories consists of a map on objects and, for each pair of objects $X, Y \in \mathcal{D}$, a morphism $P_{X,Y} : \mathcal{D}(X, Y) \rightarrow \mathcal{E}(P(X), P(Y))$ in \mathcal{V} , satisfying functoriality axioms.

Definition 2 (\mathcal{V} -Category of \mathcal{V} -Presheaves). The category of \mathcal{V} -presheaves on \mathcal{C} , denoted $[\mathcal{C}^{op}, \mathcal{V}]$, is the \mathcal{V} -category whose: ① Objects are \mathcal{V} -functors $P : \mathcal{C}^{op} \rightarrow \mathcal{V}$. ② Hom-objects $[P, Q]$ for $P, Q \in [\mathcal{C}^{op}, \mathcal{V}]$ are given by the end $\int_{X \in \mathcal{C}} \mathcal{V}(P(X), Q(X))$, where $\mathcal{V}(P(X), Q(X))$ is the internal hom in \mathcal{V} .

Lemma 1 (\mathcal{V} -Enriched Yoneda Lemma [7]). The \mathcal{V} -enriched Yoneda Lemma states that for any $A \in \mathcal{C}$ and any \mathcal{V} -presheaf $P : \mathcal{C}^{op} \rightarrow \mathcal{V}$, there is an isomorphism in \mathcal{V} :

$$[\mathcal{C}^{op}, \mathcal{V}](Y(A), P) \cong_{\mathcal{V}} P(A). \quad (1)$$

This isomorphism is \mathcal{V} -natural in A and P .

Corollary 1 (Yoneda Embedding [7, 12]). For any two objects A, B in a locally small category \mathcal{C} , there is a bijection:

$$\text{Nat}(\text{Hom}_{\mathcal{C}}(A, -), \text{Hom}_{\mathcal{C}}(B, -)) \cong \text{Hom}_{\mathcal{C}}(B, A). \quad (2)$$

This demonstrates that the functor $Y : \mathcal{C}^{op} \rightarrow [\mathcal{C}, \mathbf{Set}]$, defined by $Y(A) = h_A = \text{Hom}_{\mathcal{C}}(A, -)$, is fully faithful. This functor Y is known as the Yoneda embedding.

Theorem 1. The \mathcal{V} -enriched Yoneda embedding $Y : \mathcal{C} \rightarrow [\mathcal{C}^{op}, \mathcal{V}]$ for the sample category \mathcal{C} enriched over $\mathcal{V} = ([0, \infty], \geq, 0, +)$ with the cost function d is \mathcal{V} -fully faithful.

Proof of Theorem 1. A \mathcal{V} -functor $F : \mathcal{D} \rightarrow \mathcal{E}$ is \mathcal{V} -fully faithful if for every pair of objects $X, Z \in \mathcal{D}$, the morphism $F_{X,Z} : \mathcal{D}(X, Z) \rightarrow \mathcal{E}(F(X), F(Z))$ in \mathcal{V} is an isomorphism. We need to show that for any $A, B \in \mathcal{C}$, the map $Y_{A,B} : \mathcal{C}(A, B) \rightarrow [\mathcal{C}^{op}, \mathcal{V}](Y(A), Y(B))$ is an isomorphism in \mathcal{V} .

Considering the \mathcal{V} -Yoneda Lemma, we set $P = Y(B)$. Since $Y(B) = h_B$ is a \mathcal{V} -presheaf on \mathcal{C} , we can substitute it into the lemma:

$$[\mathcal{C}^{op}, \mathcal{V}](Y(A), Y(B)) \cong_{\mathcal{V}} Y(B)(A). \quad (3)$$

By definition of the Yoneda embedding Y , $Y(B)$ is the functor h_B such that $h_B(X) = \mathcal{C}(X, B)$ for any $X \in \mathcal{C}$. Therefore, $Y(B)(A) = h_B(A) = \mathcal{C}(A, B)$.

Substituting this back into the isomorphism derived from the Yoneda Lemma, we get:

$$[\mathcal{C}^{op}, \mathcal{V}](Y(A), Y(B)) \cong_{\mathcal{V}} \mathcal{C}(A, B). \quad (4)$$

*Equal Contribution. Corresponding author: JianglinLu@outlook.com.

The map $Y_{A,B}$ is this canonical isomorphism. In our enriching category $\mathcal{V} = ([0, \infty], \geq, 0, +)$, an object is a non-negative real number. An isomorphism $x \cong_{\mathcal{V}} y$ means there are morphisms $x \rightarrow y$ (i.e., $x \geq y$) and $y \rightarrow x$ (i.e., $y \geq x$) in \mathcal{V} . This implies $x = y$. Thus, the isomorphism in \mathcal{V} is:

$$[\mathcal{C}^{op}, \mathcal{V}](Y(A), Y(B)) = \mathcal{C}(A, B). \quad (5)$$

This means that the map $Y_{A,B} : \mathcal{C}(A, B) \rightarrow [\mathcal{C}^{op}, \mathcal{V}](Y(A), Y(B))$ is an equality of non-negative real numbers. An equality is trivially an isomorphism in \mathcal{V} (since $x \geq x$ and $x \leq x$). Since $Y_{A,B}$ is an isomorphism in \mathcal{V} for all $A, B \in \mathcal{C}$, the \mathcal{V} -functor Y is \mathcal{V} -fully faithful. \square

Proposition 1. *If two samples $X_i, X_j \in \text{Ob}(\mathcal{C})$ have \mathcal{V} -naturally isomorphic Indra representations and the cost function d satisfies the identity property, then $X_i = X_j$.*

Proof of Proposition 1. Assume that $\mathcal{C}(X_i, -) \cong_{\mathcal{V}} \mathcal{C}(X_j, -)$. This implies the existence of a \mathcal{V} -natural isomorphism $\alpha : \mathcal{C}(X_i, -) \rightarrow \mathcal{C}(X_j, -)$. By the definition of \mathcal{V} -natural isomorphism, for every object $X_k \in \mathcal{C}$, there is an isomorphism $\alpha_{X_k} : \mathcal{C}(X_i, X_k) \rightarrow \mathcal{C}(X_j, X_k)$ in \mathcal{V} . In the **Cost**-category \mathcal{V} , morphisms are costs, and the order is given by \geq . An isomorphism $x \cong_{\mathcal{V}} y$ exists if and only if $x \geq y$ and $y \geq x$, which implies $x = y$. Therefore, α_{X_k} being an isomorphism means that for all $X_k \in \mathcal{C}$, $\mathcal{C}(X_i, X_k) = \mathcal{C}(X_j, X_k)$, which translates to $d(X_i, X_k) = d(X_j, X_k)$. Now, consider specific choices for X_k . Let $X_k = X_i$, then $d(X_i, X_i) = 0$ and thus $d(X_j, X_i) = 0$. Let $X_k = X_j$, then $d(X_j, X_j) = 0$ and thus $d(X_i, X_j) = 0$. By the identity property of the cost function d , $d(X_j, X_i) = 0$ implies $X_j = X_i$, and $d(X_i, X_j) = 0$ implies $X_i = X_j$. Therefore, $X_i = X_j$. \square

Theorem 2. *For any \mathcal{V} -functor $P : \mathcal{C} \rightarrow \mathcal{V}$, the \mathcal{V} -hom-object of \mathcal{V} -natural transformations from the Indra representation of sample X_i to P , denoted by $[\mathcal{C}, \mathcal{V}](\mathcal{C}(X_i, -), P)$, is \mathcal{V} -isomorphic to $P(X_i)$.*

Proof of Theorem 2. This theorem is a direct consequence of the covariant \mathcal{V} -enriched Yoneda Lemma. The lemma states that for any \mathcal{V} -category \mathcal{C} , object $A \in \mathcal{C}$, and \mathcal{V} -functor $F : \mathcal{C} \rightarrow \mathcal{V}$, there is a \mathcal{V} -natural isomorphism:

$$[\mathcal{C}, \mathcal{V}](\mathcal{C}(A, -), F) \cong_{\mathcal{V}} F(A). \quad (6)$$

Setting $A = X_i$ and $F = P$, we have:

$$[\mathcal{C}, \mathcal{V}](\mathcal{C}(X_i, -), P) \cong_{\mathcal{V}} P(X_i). \quad (7)$$

In the **Cost**-category \mathcal{V} , this isomorphism is an equality. Therefore:

$$[\mathcal{C}, \mathcal{V}](\mathcal{C}(X_i, -), P) = P(X_i). \quad (8)$$

This equality establishes a bijection between the set of \mathcal{V} -natural transformations $\alpha : \mathcal{C}(X_i, -) \rightarrow P$ and the elements of $P(X_i)$. Specifically, this bijection maps a \mathcal{V} -natural transformation α to its component at X_i evaluated at the identity morphism, $\alpha_{X_i}(id_{X_i}) \in P(X_i)$. This result demonstrates that the Indra representation $\mathcal{C}(X_i, -)$ completely captures all information about how outgoing distances from X_i can be mapped into the values of any \mathcal{V} -functor P . The structure of any such mapping is fully determined by the value $P(X_i)$. \square

Corollary 2. *The relational structure among objects in the sample category \mathcal{C} is preserved and reflected in the relationships between their Indra representations.*

Proof of Corollary 2. Let \mathcal{C} be the sample category enriched over $\mathcal{V} = ([0, \infty], \geq, 0, +)$. To demonstrate this theorem, we need to show that for any two samples $X_i, X_j \in \text{Ob}(\mathcal{C})$, the mapping:

$$Y_{X_i, X_j} : \mathcal{C}(X_i, X_j) \rightarrow [\mathcal{C}, \mathcal{V}](\mathcal{C}(X_j, -), \mathcal{C}(X_i, -)), \quad (9)$$

is a bijection. This mapping takes a morphism $f : X_i \rightarrow X_j$ to a \mathcal{V} -natural transformation $Y(f)$. According to the covariant \mathcal{V} -enriched Yoneda Lemma, we know that:

$$[\mathcal{C}, \mathcal{V}](\mathcal{C}(X_i, -), \mathcal{C}(X_j, -)) \cong_{\mathcal{V}} \mathcal{C}(X_j, X_i). \quad (10)$$

This isomorphism shows that there is a bijective correspondence between \mathcal{V} -natural transformations from $\mathcal{C}(X_i, -)$ to $\mathcal{C}(X_j, -)$ and morphisms in $\mathcal{C}(X_j, X_i)$. Since the Yoneda embedding maps morphisms of \mathcal{C} into these natural transformations, and the Yoneda Lemma proves that the set of such natural transformations is in bijection with the morphisms of \mathcal{C} , we conclude that the Indra representation preserves the relational structure of \mathcal{C} . The mapping Y_{X_i, X_j} is this bijection. \square

B Experimental Details

B.1 Datasets

The following outlines the datasets used in our experiments.

- Office-Home [14]: The Office-Home dataset is a benchmark commonly used for evaluating domain adaptation and generalization methods. It comprises approximately 15,500 images categorized into 65 classes spanning everyday objects, such as "backpack," "bike," "keyboard," and "speaker." These images are collected from four distinct domains that reflect varying degrees of domain shift: Art (Ar), Clipart (Cl), Product (Pr), and Real-World (Rw). The Art domain contains artistic renditions of objects, Clipart includes cartoon-style illustrations, Product features images from e-commerce websites with clean backgrounds, and Real-World consists of natural photographs taken with a camera. The dataset is notably challenging due to the significant differences in style, texture, and lighting across these domains, making it ideal for testing both unsupervised domain adaptation and domain generalization algorithms. Introduced in the context of deep domain adaptation, the Office-Home dataset has become a standard benchmark in the field due to its diverse domain shifts and realistic object categories.
- MS-COCO [8]: The Microsoft Common Objects in Context (MS-COCO) dataset is a large-scale benchmark designed to advance object recognition, segmentation, and captioning in complex, real-world scenes. It contains over 330,000 images, with more than 200,000 images labeled with dense instance-level annotations. These annotations span 80 object categories—including people, animals, vehicles, and household items—resulting in more than 1.5 million object instances. A key feature of MS-COCO is its focus on objects in context, meaning that the images often contain multiple objects arranged in natural scenes with occlusions and varying spatial relationships. In addition to bounding boxes and segmentation masks, the dataset includes five human-annotated captions per image, making it particularly valuable for image captioning and vision-language tasks. The dataset is split into training, validation, and test sets, with the test set further divided into "test-dev" and "test-challenge" for benchmarking competitions. MS-COCO has become a cornerstone dataset in computer vision, especially for tasks involving detection, segmentation, keypoint estimation, and multimodal learning.
- NOCAPS [1]: The NOCAPS dataset is a large-scale benchmark designed to advance the field of novel object captioning—enabling models to describe objects not present in paired image-caption training data. It comprises 15,100 images sourced from the Open Images V4 validation and test sets, each annotated with 11 human-generated captions, totaling 166,100 captions. Notably, approximately 400 object classes in the test images are either absent or rarely mentioned in the COCO Captions training set, highlighting the dataset's emphasis on evaluating a model's ability to generalize to unseen objects. The training data for nocaps includes COCO image-caption pairs and Open Images' image-level labels and object bounding boxes, without additional paired captions for the novel objects. To assess model performance, NOCAPS categorizes images into three subsets: in-domain (only COCO classes), near-domain (both COCO and novel classes), and out-of-domain (only novel classes). This structure challenges models to leverage object detection data and linguistic knowledge to generate accurate captions for a broader range of visual concepts, thereby promoting research towards more generalized and context-aware image captioning systems.
- TIMIT [5]: The TIMIT Acoustic-Phonetic Continuous Speech Corpus is a foundational dataset in speech recognition research, developed by DARPA and released in 1993. It contains 6,300 phonetically rich sentences read by 630 speakers from eight major dialect regions of the United States, ensuring broad phonetic and dialectal coverage. Each speaker reads 10 sentences, of which two are identical across all speakers (SA sentences), while the remaining are either phonetically diverse or dialect-specific. The corpus features high-quality recordings at 16 kHz sampling rate, with time-aligned orthographic, phonetic, and word-level transcriptions, as well as phoneme boundary annotations. The audio data is stored in NIST SPHERE format. TIMIT's careful design makes it a benchmark for tasks like automatic speech recognition (ASR), phoneme recognition, speaker identification, and

acoustic modeling. Despite its relatively small size by modern standards, TIMIT remains widely used for low-resource speech tasks and controlled phonetic studies due to its precision and annotation richness.

B.2 Models

The following presents the details of the foundation models used in our experiments.

Vision Foundation Models:

- ViT [4]: <https://huggingface.co/google/vit-base-patch16-384>.
- Convnext [15]: <https://huggingface.co/facebook/convnext-base-224-22k>.
- Dinov2 [10]: <https://huggingface.co/facebook/dinov2-large>.

Language Foundation Models:

- BERT [3]: <https://huggingface.co/google-bert/bert-large-uncased>.
- Roberta [9]: <https://huggingface.co/sentence-transformers/all-roberta-large-v1>.

Foundation Models

- wav2vec [13]: <https://huggingface.co/facebook/wav2vec2-base>
- wavlm [2]: <https://huggingface.co/microsoft/wavlm-base>
- hubert[6]: <https://huggingface.co/facebook/hubert-base-ls960>.

Multimodal Foundation Models

- CLIP [11]: <https://huggingface.co/openai/clip-vit-large-patch14>.
- CLAP [16]: <https://huggingface.co/laion/clap-htsat-fused>.

B.3 Experimental Details and Reproducibility

Algorithm B.3 shows the pseudo codes for our algorithm on vision and language models. To support reproducibility, we have released the source code.

Algorithm 1 Cross-Modal Matching using Indra Representation

- 1: **Input:** text_inputs, img_inputs, integer k
 - 2: **Output:** matching score $score$
 - 3: Compute text features: $T \leftarrow \text{text_model}(\text{text_inputs})$
 - 4: Compute image features: $I \leftarrow \text{img_model}(\text{img_inputs})$
 - 5: Normalize features: $\hat{T} \leftarrow \text{L2Normalize}(T)$, $\hat{I} \leftarrow \text{L2Normalize}(I)$
 - 6: Build Indra representations: $\mathbf{I}^{\mathcal{U}}$ and $\mathbf{I}^{\mathcal{Q}}$ using Eq. (5)
 - 7: Compute difference matrix: $\mathbf{D} \leftarrow \text{cdist}(\mathbf{I}^{\mathcal{U}}, \mathbf{I}^{\mathcal{Q}})$
 - 8: Select top- k nearest neighbors: $(Val, Ind) \leftarrow \text{TopK}(\mathbf{D}, k, \text{smallest})$
 - 9: Gather similarity scores: $S \leftarrow \text{Gather}(\text{atten_clip}, Ind)$
 - 10: Compute mean score: $score \leftarrow \text{Mean}(S)$
-

C Broader Impact

The proposed Indra representation draws philosophical inspiration from Indra’s Net, emphasizing the interconnectedness of data points through relational structure. This perspective encourages a move away from isolated feature encodings toward more context-aware and structure-preserving representations. Such a shift may promote greater interpretability and robust generalization in machine learning systems, particularly in settings where relational consistency matters, such as multimodal learning, scientific modeling, and human-centered AI.

By explicitly modeling the relationships between entities, our method may also contribute to fairer and more inclusive models, as it could help mitigate biases arising from treating samples independently. However, we acknowledge that relational representations may also amplify spurious correlations if the underlying data is biased. Care must be taken to ensure that such structures are learned from ethically sourced and diverse datasets.

In the long term, we hope that incorporating ideas from philosophical traditions like Indra’s Net can broaden the conceptual foundation of AI, fostering interdisciplinary thinking and encouraging researchers to explore more holistic and principled approaches to representation learning.

References

- [1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.
- [2] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1*, pages 4171–4186, 2019.
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [5] John S Garofolo, Lori F Lamel, William M Fisher, David S Pallett, Nancy L Dahlgren, Victor Zue, and Jonathan G Fiscus. Timit acoustic-phonetic continuous speech corpus. (*No Title*), 1993.
- [6] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.
- [7] Gregory Maxwell Kelly. *Basic Concepts of Enriched Category Theory*, volume 64 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 1982.
- [8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.
- [9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [12] Emily Riehl. *Category theory in context*. Courier Dover Publications, 2017.

- [13] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.
- [14] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.
- [15] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [16] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.